

# Differential gene expression landscape of co-existing cervical pre-cancer lesions using RNA-seq

Kathryn E. Royse<sup>1</sup>, Degui Zhi<sup>2</sup>, Michael G. Conner<sup>3</sup>, Buffie Clodfelder-Miller<sup>4</sup>, Vinodh Srinivasasainagendra<sup>2</sup>, Laura Kelly Vaughan<sup>2</sup>, Christine F. Skibola<sup>1</sup>, David K. Crossman<sup>5</sup>, Shawn Levy<sup>6</sup> and Sadeep Shrestha<sup>1\*</sup>

<sup>1</sup> Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA

<sup>2</sup> Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA

<sup>3</sup> Department of Pathology, University of Alabama at Birmingham, Birmingham, AL, USA

<sup>4</sup> Cellular and Molecular Neuropathology Core, University of Alabama at Birmingham, Birmingham, AL, USA

<sup>5</sup> Department of Genetics, University of Alabama at Birmingham, Birmingham, AL, USA

<sup>6</sup> Hudson Alpha Institute for Biotechnology, Huntsville, AL, USA

## Edited by:

Mike Eccles, University of Otago, New Zealand

## Reviewed by:

Cristin Gregor Print, University of Auckland, New Zealand

Aniruddha Chatterjee, University of Otago, New Zealand

## \*Correspondence:

Sadeep Shrestha, Department of Epidemiology, University of Alabama at Birmingham, 1665 University Blvd, RPHB Room 217L, Birmingham, AL 35294-0022, USA  
e-mail: [sshrestha@uab.edu](mailto:sshrestha@uab.edu)

Genetic changes occurring in different stages of pre-cancer lesions reflect causal events initiating and promoting the progression to cancer. Co-existing pre-cancerous lesions including low- and high-grade squamous intraepithelial lesion (LGSIL and HGSIL), and adjacent “normal” cervical epithelium from six formalin-fixed paraffin-embedded samples were selected. Tissues from these 18 samples were isolated using laser-capture microdissection, RNA was extracted and sequenced. RNA-sequencing generated 2.4 billion raw reads in 18 samples, of which ~50.1% mapped to known and annotated genes in the human genome. There were 40 genes up-regulated and 3 down-regulated (normal to LGSIL) in at least one-third of the sample pairs (same direction and FDR  $p < 0.05$ ) including *S100A7* and *KLK6*. Previous studies have shown that *S100A7* and *KLK7* are up-regulated in several other cancers, whereas *CCL18*, *CFTR*, and *SLC6A14*, also differentially expressed in two samples, are up-regulated specifically in cervical cancer. These differentially expressed genes in normal to LGSIL progression were enriched in pathways related to epithelial cell differentiation, keratinocyte differentiation, peptidase, and extracellular activities. In progression from LGSIL to HGSIL, two genes were up-regulated and five down-regulated in at least two samples. Further investigations using co-existing samples, which account for all internal confounders, will provide insights to better understand progression of cervical pre-cancer.

**Keywords:** RNA-sequencing, gene expression, cervical dysplasia, co-existing lesions, human genome

## INTRODUCTION

Squamous cell carcinoma of the cervix results from a sequence of well delineated non-invasive pre-cancer stages. Based on the cytological findings of Bethesda Classification system (1), these pre-cancer stages are classified as low-grade squamous intraepithelial lesion (LGSIL), consisting of cytological atypia and histological cervical intraepithelial lesion (CIN) 1, and high-grade squamous intraepithelial lesion (HGSIL), consisting primarily of CIN 2–3 plus carcinoma *in situ* (2, 3). In 2014, it is estimated that 12,360 cases of invasive cervical cancer will occur in the United States; however, 1.25 million women are expected to be diagnosed with pre-cancer by a Papanicolaou (Pap) test (4). Similarly, ~600,000 cervical cancers are expected worldwide, with epidemic proportions of pre-cancers, mostly undiagnosed (5). Studies of cervical disease progression (6) suggest that lesions in ~60% of women with LGSIL will spontaneously regress, another 20–30% will persist unchanged, about 5–10% will progress to high-grade HGSIL, and only 1% will develop invasive carcinoma (6, 7). The likelihood of HGSIL regression is 33%; progression to invasion is >12% (6). Persistent high-risk HPV (HR-HPV) infection (8), high-viral load (9), and integration of HPV DNA (10) are likely markers or determinants of progression of pre-cancer lesions to

cervical cancer; however, host factors have not been thoroughly studied. While genetic alterations in cancer are common, changes found in different grades of pre-cancer lesions are more likely to reflect causal events initiating and promoting the progression to cancer, yet little is known about these genomic changes that occur.

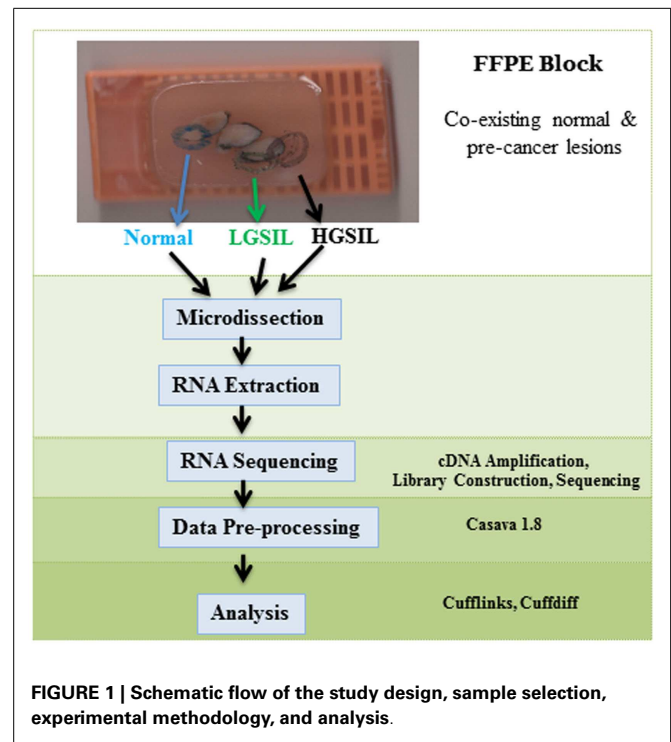
Some experts arguably consider that it is reasonable standard-care to follow low-grade pre-cancer lesions and HPV infections without active treatment. Since HPV infection and LGSIL are diagnosed in epidemic proportions, novel biomarkers with higher specificity for cervical lesions would improve cervical cancer screening. Better early detection biomarkers would also greatly assist in the stratification of patients for chemoprevention trials of pre-neoplasia. Currently, there are no validated diagnostic or prognostic biomarkers that identify LGSIL destined toward HGSIL or cervical cancer. Since HGSIL are near-obligate precursors of cervical cancer, it is standard clinical practice to use invasive surgical interventions to reduce the burden of progression to cancer. The detection of LGSIL that may progress to HGSIL and the prevention of this progression is an important and suitable goal for non-invasive medical intervention to reduce the incidence of cervical cancer.

Gene expression studies are quite sparse for HPV-related cervical dysplasia. Squamous cell carcinoma of the cervix results from a sequence of well delineated non-invasive pre-cancer stages. Quantifying the decisive physical changes, i.e., differential function and expression of genes in the co-existing “normal” cervical epithelium and neighboring pre-cancerous lesions; low-grade squamous intraepithelial lesions (LGSIL) and high-grade squamous intraepithelial lesions (HGSIL) will elucidate when, where, and to what extent genomic variations facilitate development and progression of pre-cancer. Since LGSIL and HGSIL lesions are generally related when they co-occur, these types of samples provide an opportunity to assess morphologic progression with regard to space and time, while controlling for internal confounders (11). Specifically, RNA sequencing (RNA-seq) can examine expression patterns of genes from formalin-fixed paraffin-embedded (FFPE) tissues, which can be used to study pre-cancer progression in co-existing LGSIL and HGSIL tissues (12). In the current study, we utilized laser-capture microdissection to extract specific tissues from co-existing neoplastic stages (normal, LGSIL, and HGSIL) on FFPE samples from six women who underwent loop electrosurgical extraction procedures (LEEP), and performed innovative RNA extraction and sequencing (RNA-seq) technologies to enable comprehensive gene expression profiling of selected cell types for comparison within and between individuals.

## MATERIALS AND METHODS

### STUDY POPULATION AND PATIENT SAMPLE SELECTION

The patient samples were obtained from the University of Alabama at Birmingham (UAB) Comprehensive Cancer Center (UAB-CCC) tissue procurement shared facility (TPSF) where standard protocols are followed to routinely collect cervical samples from the UAB colposcopy clinic and preserve as FFPE tissues (13). Since formalin fixation methodology, which affects nucleic acid integrity, can vary among hospital laboratories, we used only samples from women that received LEEP treatment at the UAB colposcopy clinics during June 2010 to April 2012 for abnormal cytology. A UAB pathologist prospectively reviewed cervical tissue samples from ~850 women, aged 20–25 years old to confirm if they had a co-existing spectrum of normal and pre-cancer LGSIL and HGSIL on the sample block (Figure 1). If co-existing lesions (LGSIL and HGSIL) were found for LEEP treatments during that period, women must have had HGSIL (CIN2/3) confirmed biopsies within 6 months of their treatment. Samples with evidence of immunosuppression or HIV infection were excluded due to known differences in rates of cervical abnormalities and pathogenesis (14, 15). Following these criteria, we identified 10 FFPE blocks from European American women (Figure S1 in Supplementary Material) of whom 6 were similar in relation to information on referral, cytology, demographic characteristics, and subsequent histologic biopsy (Table 1) are used for the study. Specific laboratory methods (Figure 1) were followed as specified below. The study protocols using these samples conformed to human-experimentation guidelines set forth by the United States Department of Health and Human Services and was reviewed and approved by institutional review board (IRB) at UAB.



### SLIDE PREPARATION

Formalin-fixed paraffin-embedded blocks from six participants were used for slide preparation and microdissection. First, the pathologist, an expert in gynecologic lesion analysis, identified and marked 12 pre-cancerous lesions (LGSIL and HGSIL in each) and 6 co-existing normal tissues, respectively, to ensure the correct cell populations. Then, depending on the size of the tissues on the block, two to six 10 µm sections from the FFPE blocks were cut using RNase-free blades at each area identified by the pathologist and mounted on the coated slide (Leica Microsystems, Buffalo Grove, IL, USA). To prevent contamination, a new blade and a different slide was prepared for each cell type (normal, LGSIL, and HGSIL) and captured separately. The samples were heated in a water bath (58°) for 2 min and the paraffin debris was carefully scraped off. No staining of samples was performed since it can further degrade RNA in FFPE tissues.

### MICRODISSECTION METHOD

To reduce sample degradation, all laboratory procedures including sample preparation, microdissection, and RNA extraction were performed within a 24 h period. Microdissection of the 18 specific tissues (adjacent normal, LGSIL, and HGSIL) from 6 participants was performed to reduce possible contamination of different cell types in heterogeneous tissue, which could mask the detection of gene expression alterations in the cells of interest. Slide preparation, laser-capture microdissection (LCM), and RNA extraction was conducted at LifeSpan Molecular Pathology Core Services lab in Providence, Rhode Island. LCM was conducted using the Arcurus Veritas Laser-Capture Microdissection (Mountain View, CA, USA) to isolate individual cell populations from distinct normal and pre-cancer LGSIL and HGSIL cells, which were then placed

**Table 1 | Demographic and LEEP diagnosis information from study of cervical dysplasia FFPE co-existing HGSIL and LGSIL lesions.**

Sample ID number	Age	Smoker	HGSIL LEEP diagnosis	ICD9 code	Keratinizing dysplasia	Glandular involvement or extension	Loop margins	Adjacent normal sample location
#109	25	No	CIS of cervix uteri (includes all CIN3)	233.1	Yes	Yes	—	HGSIL
#110	23	Yes	CIS of cervix uteri (includes all CIN3)	233.1	No	No	—	HGSIL
#111	24	Yes	CIS of cervix uteri (includes all CIN3)	233.1	No	No	+	—
#113	29	Yes	Moderate dysplasia of cervix (CIN2)	622.12	No	No	+	LGSIL
#115	25	Unknown	Moderate dysplasia of cervix (CIN2)	622.12	Yes	Yes	+	LGSIL
#116	23	Unknown	CIS of cervix uteri (includes all CIN3)	233.1	No	No	—	LGSIL

on CapSure HS LCM Caps using the UV capture laser protocols. Adjacent stained slides were used as a roadmap to determine which areas would be collected. The relatively low intensity of the laser does not damage DNA, RNA, or proteins in the captured cells (16). Approximately 10,000 cells were collected per tissue sample onto a single LCM cap. Each population of captured cells was re-examined by microscopic visualization for confirmation, and before and after images of each sample were taken (Figure S1 in Supplementary Material) for additional confirmation by a gynecologic pathologist specialized in cancers.

### RNA EXTRACTION

Total RNA was extracted from all 18 microdissected tissues from 6 to 10 LCM using the Qiagen AllPrep DNA/RNA Micro Kit (Qiagen, CA, USA), according to the manufacturer's instructions, with extended proteinase K digestion. Prior to amplification, the quality and level of degradation of the extracted RNA was assessed with RIN (RNA integrity number) assigned by the Agilent 2100 Bioanalyzer instrument using the RNA 6000 Pico kit (Agilent Technologies, Santa Clara, CA, USA). All 18 samples were stored in nuclease-free tubes and stored at  $-80^{\circ}\text{C}$  until shipment for analysis.

### RNA PRE-PROCESSING, LIBRARY PREPARATION, AND SEQUENCING

To increase low-RNA yields, the Nugen Ovation © FFPE RNA-Seq System (Nugen, CA, USA) was used to amplify cDNA from total RNA for transcriptome sequencing. Amplification was initiated at the 3' end and also randomly throughout the whole transcriptome in the sample. As described in user's manual, amplification of both mRNA and non-polyadenylated transcripts makes the Ovation RNA-Seq System (Nugen, San Carlos, CA, USA) ideal for amplification prior to next generation sequencing (NGS). The amplified cDNA was quantified using Qubit dsDNA BR Assay 2.0 (Invitrogen Life Technologies, Grand Island, NY, USA) and all samples produced ample cDNA yields (1.26–1.89 ng/ $\mu\text{l}$ ) for library construction. Then 1  $\mu\text{l}$  from each sample was diluted 5 $\times$  for analysis by Agilent DNA1000 chip to verify that all profiles obtained fit NuGen's expected profile. None of the samples were sonicated because QC analysis done after the Ovation RNAseq FFPE protocol showed that the average fragment sizes (417–466 bp) were in an acceptable range for library prep. Sequencing libraries for whole transcriptome analysis were prepared with 2  $\mu\text{g}$  of each sample using Illumina Tru-Seq © Library Preparation Kits (Illumina, San Diego, CA, USA). Two 7-base index sequences were used to prepare bar-coded libraries for sample multiplexing. Two indexed libraries

were loaded into each lane of flow cells. Sequencing was performed using Illumina HiSeqH2000 (Illumina, San Diego, CA, USA) at 50 base pairs, paired end, and 25 million paired reads per sample following the manufacturer's protocol. Multiplexed single-read runs were carried out with a total of 107 cycles. The RNA-Seq dataset has been deposited in the NCBI SRA with accession ID SRP048735.

### DATA PRE-PROCESSING AND ALIGNMENT OF SEQUENCED READS

FASTQ sequence files were generated from the raw base-call data for all 18 samples using CASAVA 1.8 (Illumina, CA, USA), the standard data processing package from Illumina. De-multiplexing of sample indices was set with one mismatch tolerance to separate the two samples within each lane. Quality assessment of raw FASTQ reads was performed using the FASTQC program, as previously described (17).

The paired end reads were aligned against the Ensembl GRCh37.62 B (hg19) reference genome using 150 as the mate pair means inner distance and the pre-set settings in TopHat v2.0.6. The alignment quality and distribution of the reads were estimated using SAMTools v1.18 (Illumina, CA, USA). The reference genome-guided transcript assembly of the aligned reads was performed using quartile normalization, bias correction, and the default assembly settings for Cufflinks v2.1.1. The transcript abundance was calculated by estimating the fragments per kilobase of exon per million mapped fragments (FPKM), and all expressed transcripts were binned (0, <1, 1–10, and >10) on the basis of their abundance (FPKM).

### DATA VISUALIZATION AND CLUSTERING ANALYSIS

The similarity of the relative gene transcript abundances (using Log2 transformed values of FPKM) for each of the 18 samples were compared using Spearman's correlation calculation in SAS and an unsupervised hierarchical clustering analysis in Gene-E, a visualization tool developed by the Broad Institute (<http://www.broadinstitute.org/cancer/software/GENE-E/>). The pairwise comparisons of FPKM between normal to LGSIL and LGSIL to HGSIL in each sample were compared between the tissues from these three stages and visualized as stacked histograms by genomic location in each chromosome using Circos plots (18).

### DIFFERENTIAL GENE EXPRESSION ANALYSIS

UCSC known genes annotation was used to assess differentially expressed genes for the pre-cancer stage effect using CuffDiff tools (19) in two different methods: (1) grouping all six samples by

their tissue classification and comparing the geometric averages progressively between normal vs. LGSIL and LGSIL vs. HGSIL; and (2) comparing the geometric averages progressively within each individual separately (six pairs for normal to LGSIL and six pairs for LGSIL to HGSIL) using CuffDiff, and then merging the results across all individuals. Because of the small sample size and unique sample type, both methods were assessed to fully capture the variability in the structured experimental design. To account and adjust for multiple testing, the FDR  $q$ -values were calculated by CuffDiff from the raw  $p$ -values, which are estimated using beta negative binomial tests of variance in read counts (19). Transcripts were considered to be differentially expressed if their expression values (log2) differed by a factor of 1.5 and FDR < 0.05. For FPKM = 0, zero to positive KPM was considered positive infinity and positive KPM to zero was considered negative infinity (Tables S3A,B in Supplementary Material). Volcano plots of  $-\log_{10}(p\text{-value})$  vs. Log2 (FPKM) fold-change were made to examine these association in each tissue pair within each individual.

### PATHWAY ENRICHMENT ANALYSIS

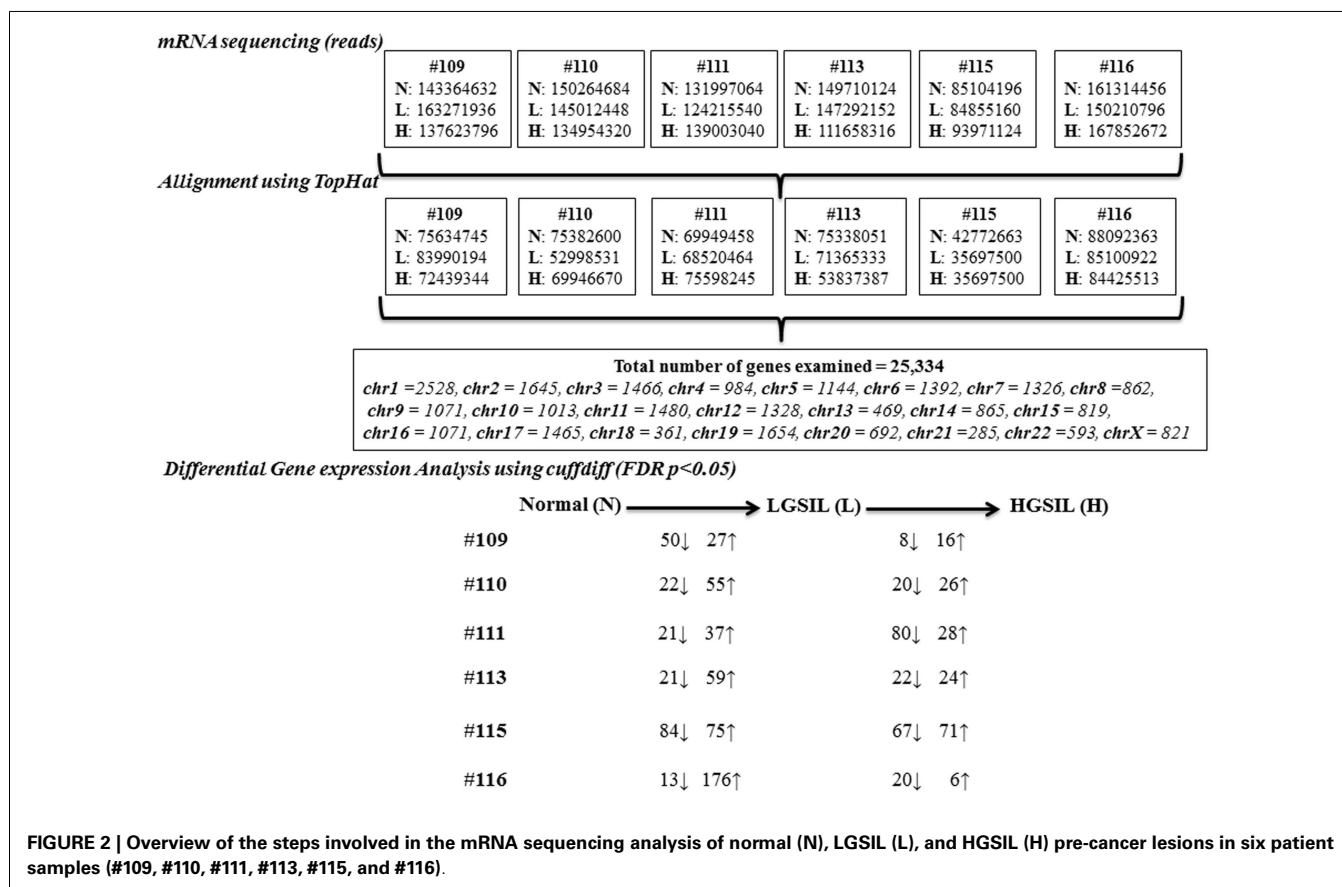
WEB-based GENE SeT AnaLysis Toolkit (WebGestalt) was used for gene ontology (GO) analysis to identify pathways that were enriched in all significant gene lists by each of the progression stage pairs (LGSIL vs. Normal, HGSIL vs. LGSIL) (20). Only genes that were differentially expressed in at least two participant sample comparisons (both in direction and significance at FDR  $q \leq 0.05$ )

were included in the analysis. Statistical significance was estimated using the hypergeometric test and GO categories were considered to be significant if the pathway included at least five genes and the adjusted  $p$ -value of enrichment using a Benjamini–Hochberg FDR correction was  $\leq 0.05$ .

### RESULTS

The mRNA sequencing of the co-existing spectrum of normal and pre-cancer LGSIL and HGSIL on the sample block from 6 patients (18 samples) generated a total 2.4 billion raw reads, ranging from 264 to 470 million reads per person (Figure 2; Table S1 in Supplementary Material). The overall raw read mean quality score was high (mean Phred Quality Score = 36.76) with 93.5% of bases above Q30. Among the 2.4 billion high-quality raw reads, 50.1% of the reads were mapped to the human genome with known gene annotations. Of the 29,061 genes across 22 autosomal and X chromosomes, 3727 of them had zero FPKM in all samples so 25,334 genes were examined for differential gene expression (Table S2 in Supplementary Material).

As shown in the unsupervised hierarchical clustering dendrogram (Figure S2A in Supplementary Material), most of the samples from each cervical pre-cancer stage clustered together by sample and then by stage in our analysis. Sample #111 showed the most differences within the three stages and sample #115, which clustered all three stages, seemed to be in a different branch. Overall, more genes were up-regulated during progression from normal to LGSIL, with samples #115 and #116 having the largest numbers



(Figure 2; Table S3A in Supplementary Material). On the other hand, more genes were down-regulated during progression from LGSIL to HGSIL, with samples #111 and #115 having the largest number of genes (Figure 2; Table S3B in Supplementary Material). Down-regulation of genes is a hallmark signature of advanced pre-cancer progression to cancer. One patient sample (#115), which consistently had a larger number of differentially expressed genes in both progression from normal to LGSIL and LGSIL to HGSIL, was one of two samples that displayed moderate dysplasia (CIN2).

None of the 25,334 genes were differentially expressed after stringent statistical correction (FDR  $p \leq 0.05$ ), using the first method when the analysis performed was combined for all merged samples (by pre-cancer stage) (Tables S3A,B in Supplementary Material). However, three genes (*C1orf120*, *EPGN*, and *NDRG4*) were differentially expressed (FDR  $q \leq 0.20$ ) during progression from normal to LGSIL and six genes (*ADH7*, *AX746562*, *CRNN*, *FABP*, *SLN*, and *SPRR3*) were differentially expressed (FDR  $q \leq 0.20$ ) during progression from LGSIL to HGSIL as listed in Table 2 (all results in Tables S3A,B in Supplementary Material).

The main analysis using the second method was focused on the differentially expressed genes (FDR  $p \leq 0.05$ ) identified by pairwise comparison of each sample separately using CuffDiff (Figure S3 in Supplementary Material). Between 6 patient samples of normal to LGSIL pairs, there were a total of 211 genes down-regulated and 429 up-regulated (Figure 3A). Of these, *S100A7* and *KLK6* were differentially up-regulated in 3 of 6 patient samples (FDR  $q \leq 0.05$ ); whereas, *CCL18*, *FLT3*, and *RORC* were down-regulated and 38 genes were up-regulated in 2 of 6 patients (Figure 3B). In progression to high-grade pre-cancer (LGSIL to HGSIL), although there were a total of 187 different genes up-regulated in various samples, the majority (165) were present in one patient sample (#115), which displayed moderate dysplasia. Overall, *C12orf63*, *KRTDAP*, *SBSN*, *FABP4*, *CBLN1* were down-regulated whereas *CTCF* and *PLAC8L1* were up-regulated in at least two of six samples (Table 3). When comparing the progression from LGSIL to HGSIL in these six patient samples, seven genes were differentially expressed in at least two samples (Table 3).

Of the 43 differentially expressed genes in at least 2 participant samples in progression from normal to LGSIL (Table 3), 7 (*CLCA2*, *CRCT1*, *PKP1*, *S100A7*, *PLCH2*, *SPRR1B*, and *NADK*)

were mapped to chromosome 1 and another 7 (*KLK6*, *CASP14*, *CNFN*, *KLK12*, *KLK11*, *KLK10*, *KLK13*) to chromosome 19. Cytoband location 19q13.3–q13.4 was common for differentially expressed *KLK* gene family (*KLK12*, *KLK11*, *KLK10*, and *KLK13*), the serine proteases encoded from this Kallikrein gene family have been implicated in various cancers (21). Of the 7 genes differentially expressed for LGSIL to HGSIL in at least 2/6 individuals (Table 3), *KRTDAP* and *SBSN* were mapped to chromosome 19 with the others mapped to different chromosomes.

With progression from normal to LGSIL, six genes (*SPRR1B*, *CNFN*, *TP63*, *PAX2*, *SCEL*, *S100A7*) were enriched in cancer pathways including GO:0030855, epithelial cell differentiation (adj  $p \leq 7.4 \times 10^{-3}$ ), and GO:0008544, epidermis development (adj  $p = 0.05$ ). Five genes (*CNFN*, *SPRR1B*, *S100A7*, *SCEL*, *TP63*) were enriched in GO:0030216 ~ keratinocyte differentiation ( $p \leq 1.10 \times 10^{-3}$ ) and GO:0009913, epidermal cell differentiation (adj  $p \leq 1.3 \times 10^{-3}$ ). Several other pathways (Figure S4A in Supplementary Material) were enriched with the clusters of differentially expressed genes: GO:0017171, serine hydrolase activity (5 genes, adj  $p \leq 1.0 \times 10^{-3}$ ), GO:0004252, serine-type endopeptidase activity (5 genes, adj  $p \leq 1.0 \times 10^{-3}$ ), GO:0004175, endopeptidase activity (6 genes, adj  $p \leq 1.8 \times 10^{-3}$ ), GO:0008509 anion transmembrane transporter activity (5 genes, adj  $p \leq 1.8 \times 10^{-3}$ ), GO:0070011 peptidase activity, acting on L-amino acid peptides (6 genes, adj  $p \leq 1.32 \times 10^{-2}$ ), GO:0008233 peptidase activity (6 genes, adj  $p \leq 1.51 \times 10^{-2}$ ), and GO:0005576, extracellular region (13 genes, adj  $p \leq 1.82 \times 10^{-2}$ ). No distinct pathways (with more than five genes and adj  $p \leq 0.05$ ) were enriched with genes in at least two pairs of patient samples in progression from LGSIL to HGSIL (Figure S4B in Supplementary Material).

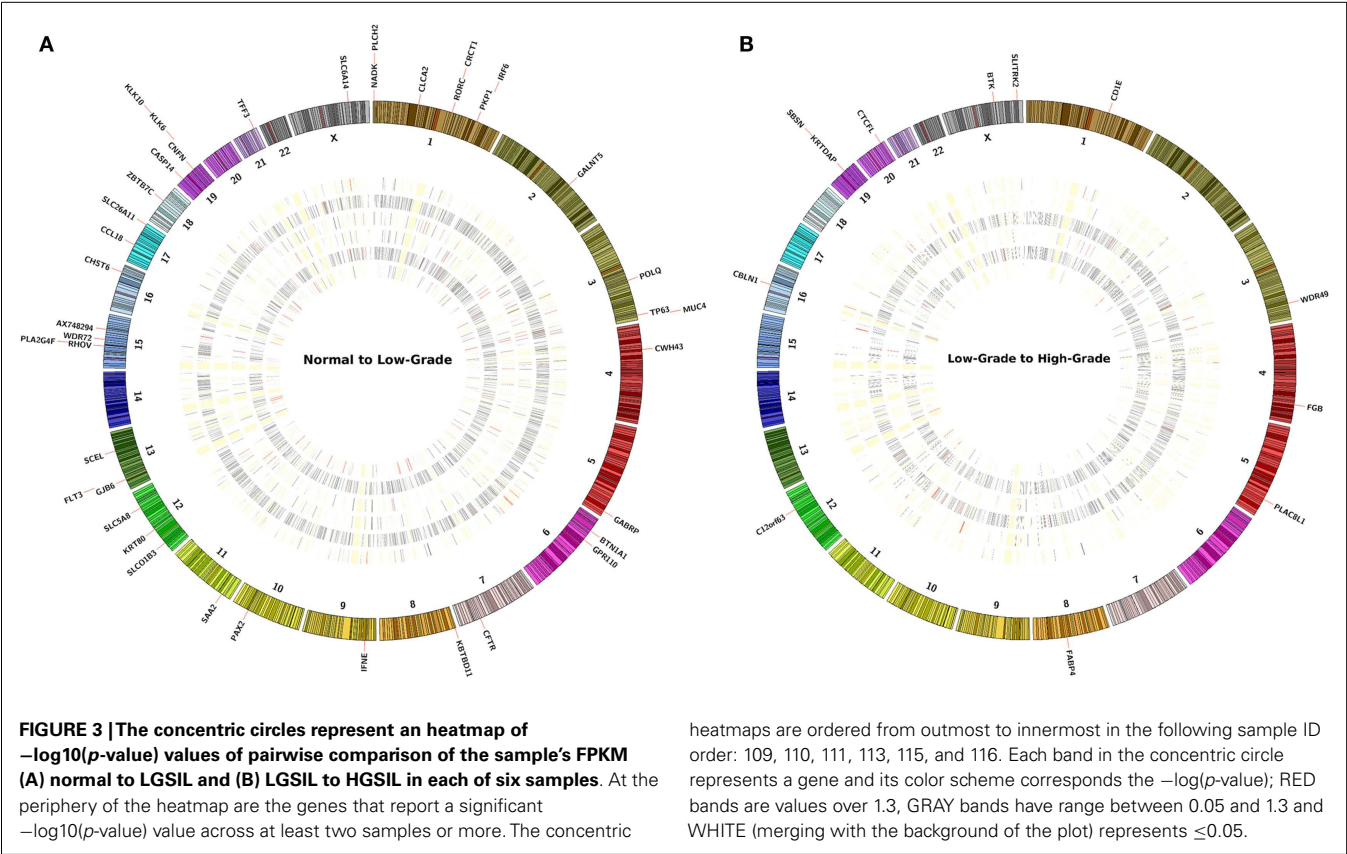
## DISCUSSION

We used next generation RNA-seq methods to study the transcriptomic landscapes of co-existing pre-cancer lesions and understand the mechanism of pre-cancer progression from normal to LGSIL and LGSIL to HGSIL in six individuals. To our knowledge, there have been no published reports of assessment of transcriptomic expression patterns from co-existing spectrum of neoplasia from a single sample. In most studies, samples from different individuals or from different times from the same individual are compared. A

**Table 2 | Differentially expressed genes (FDR  $q \leq 0.20$ ) from transcriptome analysis of six samples combined – normal to LGSIL and LGSIL to HGSIL (Log2 transformed).**

Gene	Locus	Normal	LGSIL	Fold-change	p-Value	q-Value
<b>Normal to LGSIL</b>						
<i>C1orf120</i>	chr1:182376755-182383948	11.0013	0	–inf	5.00E-05	0.18
<i>EPGN</i>	chr4:74979895-75196255	0	3.33617	+inf	5.00E-05	0.18
<i>NDRG4</i>	chr16:58497548-58547522	7.04797	85.1306	3.5944	5.00E-05	0.18
<b>LGSIL to HGSIL</b>						
<i>ADH7</i>	chr4:100333417-100356525	2.89046	0	–inf	5.00E-05	0.18
<i>AX746562</i>	chr7:44406738-44408891	0	4.34595	inf	5.00E-05	0.18
<i>CRNN</i>	chr1:152381718-152386750	194.411	2.41024	–6.33379	5.00E-05	0.18
<i>FABP4</i>	chr8:82390731-82395473	237.894	15.9085	–3.90245	5.00E-05	0.18
<i>SLN</i>	chr11:107578108-107582787	0	25.907	inf	5.00E-05	0.18
<i>SPRR3</i>	chr1:152974222-152976332	350.506	14.4714	–4.59817	5.00E-05	0.18





**Table 3 |** Differentially expressed genes in 6 pairs of normal-LGSIL-HGSIL ( $\log_2$  fold-change  $> 1.5$  and  $FDR\ p \leq 0.05$ ).

Common pairs	Normal-to-LGSIL ( <i>q</i> -value)	LGSIL-to-HGSIL ( <i>q</i> -value)
3/6 Pairs	<b>All up-regulated</b> <i>KLK6, S100A7</i> <b>2 up-regulated and 1 down-regulated</b> <i>ZBTB7C, GALNT5</i>	
2/6 Pairs	<b>Up-regulated</b> <i>AK026736, AX748294, BTN1A1, CASP14, CFTR, CHST6, CKMT1A, CLCA2, CNFN, CRCT1, CWH43, DQ586546, GJB6, HCG22, IFNE, KBTBD11, KLK10, KLK11, KLK12, KLK13, KRT80, MUC21, NADK, PAX2, PKP1, PLA2G4F, PLCH2, POLQ, RHOV, SAA2, SCEL, SLC26A11, SLC5A8, SLC6A14, SLC01B3, SPRR1B, TFF3, TP63</i> <b>1 up-regulated and 1 down-regulated</b> <i>GABRP, GPR110, IRF6, LY6G6C, MUC4, WDR72</i> <b>Down-regulated</b> <i>CCL18, FLT3, RORC</i>	<b>Up-regulated</b> <i>CTCFL, PLAC8L1</i>  <b>1 up-regulated and 1 down-regulated</b> <i>BTK, C9orf122, CD1E, FGB, SLITRK2, WDR49</i> <b>Down-regulated</b> <i>C12orf63, KRTDAP, SBSN, CBLN1, FABP4</i>

major advance in functional genomic investigations is the use of NGS, RNA-seq with *ex vivo*-derived genetic material originating from morphologically distinct cellular subpopulations within tissue. The first application combining LCM and cDNA microarray technologies to analyze gene expression in breast cancer specimens was reported in 1999 (22). Since then, most investigations of changes in gene expression associated with the progression stages

in cancer have been targeted for breast cancer, prostate cancer, lung cancer, and colorectal cancer (23–30). Reports of gene expression studies of cervical cancer are sparse, specifically for pre-cancer lesions (31–33). Using gene expression transcripts clustered by pre-cancer lesion grades, several genes were either down- or up-regulated during the progression process in each individual (Figures 3A,B). However,

43 genes were differentially expressed (in the same direction and statistically significant) in the samples from normal to LGSIL and 7 genes were differentially expressed from LGSIL to HGSIL in 2 or more individuals of the 6 participants (Table 3). In particular, *S100A7* and *KLK6* genes were differentially expressed during progression from normal tissue to LGSIL in 3/6 samples and could be key members of signature networks in cervical pre-cancer progression. The *KLK6* gene was differentially up-regulated in 3/6 LGSIL samples compared to HGSIL; whereas, *KLK10*, *11*, *12*, and *13* were also up-regulated in 2/6 samples. The *KLK* gene family is a member of the protease clan PA, protease family S1 with subfamily A, located on chromosome 19q13.3–q13.4. *KLK6* along with *10* and *11* and *13* are emerging biomarkers for ovarian and cancer (34). Moreover, *KLK6* has been differentially expressed in breast, uterine, and colon cancers (35–37). The *S100A7* gene, also known as psoriasin, is located within the epidermal differentiation complex on human chromosome 1q21 (38) and plays an important role as an immunomodulatory protein in skin (39). Increased *S100A7* expression has been reported in several epithelial malignancies, including head and neck squamous cell carcinoma (HNSCC) and oral dysplasia (40), which shares many features, including HPV infection, with cervical SCC, as well as skin, bladder, breast cancer, and adenocarcinomas of the stomach (41). However, we did not test for HPV and could not assess the interactions of these genes with HPV.

Three other genes are also noteworthy, chemokine (C–C motif) ligand 18 (*CCL18*) (42), cystic fibrosis transmembrane conductance regulator (*CFTR*) (43), and solute carrier family 6 member 14 (*SLC6A14*), as they have been reported to be up-regulated in cervical cancer malignancy and in the same direction as in progression from normal to LGSIL in our study (44). *CCL18* is known to be involved in adaptive immune system and its role in cancer is not fully known but is best known for inducing metastasis of breast cancer cells by binding to PITPNM3 (45). *CFTR*, found in epithelial cells functions as a cAMP-activated ATP-gated anion channel and *SLC6A14* functions as Na<sup>+</sup> and Cl<sup>–</sup> dependent neutral and basic amino acid transporter, but their roles in cancer are not known. Of note, two genes *ZBTB7C* and *GALNT5* were differentially expressed in the same direction in two samples but opposite in the third in normal to LGSIL progression. The pathway analysis of these genes provided further insight into the expression profiling of genes involved in pre-cancer progression. The functions of the genes identified in the progression from normal to LGSIL were enriched in the biological processes and pathways of keratinocyte differentiation, epidermis development, peptidase activities, and extracellular region activities.

Inclusion of pre-cancer tissues from the same individual with respect to time and space is a major strength of this approach where all internal confounders are controlled for in determining signature gene expression patterns. However, there might be residual heterogeneity in samples between patients (Figure S2 in Supplementary Material); thus, the analysis from progression within each individual might be more informative than the combined analysis taking the mean or median values across different stages from different individuals. For instance, four patient samples had ICD9 code 233.1 (carcinoma *in situ*) (Table 1) and three of them (samples #109, #110, and #116) had *S100A7* genes

up-regulated and another set (samples #110, #111, and #116) had *KLK6* up-regulated in normal to LGSIL pairs. Samples #113 and #115 may have differing cellular makeups since their HGSIL were diagnosed as moderate dysplasia (CIN2) (ICD9 code 622.12). Microdissection methods create a homogeneous sample, but any error in procedures can include adjacent heterogeneous tissues. For example, in samples #109 and #110, HGSIL cells were adjacent to normal samples (Table 1) and contaminated samples can also misclassify the expression patterns. Also, microarray technologies typically require fresh tissues; however, FFPE tissues have been regularly used in clinical research. With careful selection of samples, such as the age of the sample, length, and conditions of storage (13, 46), gene expression analyses of FFPE tissues have shown comparable results to frozen tissues (47–49). In addition, various new technologies have proven that RNA acquired by laser-capture microdissection from FFPE samples also yields reliable microarray and NGS data (50). Our use of LCM to separate a heterogeneous sample as well as presence of common gene signatures for cervical pre-cancer lead us to have more confidence in the phenotype and therefore results. Utilization of FFPE blocks in such genomic approaches is innovative as they can provide tremendous resources and opportunities for epidemiological, basic, and translational studies.

While effective prophylactic vaccines are available (for HPV types 6, 11, 16, and 18), there is no treatment for infected individuals and logistic issues of vaccine delivery still exist. Thus, discovering biomarkers that are associated with the progression of pre-cancer lesions and cancer are extremely important. We are cognizant that our study is limited by number of patient samples and our findings need to be replicated in larger study. Thus, it is unknown if some of these gene patterns observed might be rare and specific to individual patients and thus may not be used as universal biomarker for progression. However, our approach of using co-existing samples is novel and may enable accurate aggregate genomic information regarding the potential mechanism or pathway of progression, since all known and unknown confounding factors are controlled for by internal comparisons. The identified differentially expressed genes (validated in two or three samples), specifically *S100A7*, *KLK6*, *CCL18*, *CFTR*, and *SLC6A14*, with reports of involvement in other cancers or novel ones with potentially related biological pathways, represent targets for understanding mechanisms of pre-cancer progression.

Since complex processes regulate gene expression, it is very likely that not all important transcriptome genes were identified and conversely that some gene expression changes identified in this study may not later be confirmed, and therefore, the results of this study should be replicated in larger studies. Additionally, we chose to only assess differential gene expression in co-existing lesions although there are other processes and markers of expression including RNA-splicing and transcript expression that were not assessed in the analysis. Despite these limitations, we successfully designed and analyzed a novel application of RNA-Seq technology to identify and annotate genes and networks that may be present in cervical co-existing lesions. Future studies can examine other complexities of the transcriptome using the RNA-seq data including splice junctions, fusion, allelic variants including somatic mutations, and HPV integration that could aggregately either

explain these differential gene expression patterns or progression of pre-cancer.

## AUTHOR CONTRIBUTIONS

Sadeep Shrestha directed all aspects of the cervical pre-cancer transcriptome project. Kathryn E. Royse and Sadeep Shrestha designed the experiments and wrote the main manuscript text with help from Degui Zhi and Christine F. Skibola. Degui Zhi led the overall data QC and analyses with assistance from Kathryn E. Royse, Vinodh Srinivasasainagendra, Laura Kelly Vaughan, and David K. Crossman. Michael G. Conner helped identify the samples and pre-cancer lesions, Buffie Clodfelder-Miller provided technical assistance with microdissection protocol, and Shawn Levy directed the RNA sequencing experiments.

## ACKNOWLEDGMENTS

We thank Dr. Elizabeth E. Brown, Dr. Xiangqin Cui, Dr. John Waterbor, and Dr. Ronald D. Alvarez for their insights with study design, analysis, and interpretation of the data. RNA sequencing and laboratory methods were supported by the Pilot Project Award (PI: Sadeep Shrestha) from the Johns Hopkins Cervical Cancer SPORE (2P50CA098252 PI: T. C. Wu). Kathryn E. Royse was supported in part by the NIH Cancer Prevention and Control Training Program (R25CA47888). Computational portions of this research were supported by NIH S10RR026723. This work was supported in part by the research computing resources acquired and managed by UAB IT Research Computing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at <http://www.frontiersin.org/Journal/10.3389/fonc.2014.00339/abstract>

## REFERENCES

- Luff RD. The Bethesda system for reporting cervical/vaginal cytologic diagnoses: report of the 1991 Bethesda workshop. *Hum Pathol* (1992) **23**:719. doi:10.1016/0046-8177(92)90338-4
- Davey DD, Woodhouse S, Styer P, Stastny J, Mody D. Atypical epithelial cells and specimen adequacy: current laboratory practices of participants in the college of American pathologists interlaboratory comparison program in cervico-vaginal cytology. *Arch Pathol Lab Med* (2000) **124**:203–11. doi:10.1043/0003-9985(2000)124<0203:AECASA>2.0.CO;2
- Schiffman MH, Castle P. Epidemiologic studies of a necessary causal risk factor: human papillomavirus infection and cervical neoplasia. *J Natl Cancer Inst* (2003) **95**:E2. doi:10.1093/jnci/95.6.E2
- American Cancer Society. *Cancer Facts & Figures*. Atlanta, GA: American Cancer Society (2014).
- Arbyn M, De Sanjose S, Saraiya M, Sideri M, Palefsky J, Lacey C, et al. EUROGIN 2011 roadmap on prevention and treatment of HPV-related disease. *Int J Cancer* (2012) **131**:1969–82. doi:10.1002/ijc.27650
- Ostor AG. Natural history of cervical intraepithelial neoplasia: a critical review. *Int J Gynecol Pathol* (1993) **12**:186–92. doi:10.1097/00004347-199304000-00018
- Strickler HD, Burk RD, Fazzari M, Anastos K, Minkoff H, Massad LS, et al. Natural history and possible reactivation of human papillomavirus in human immunodeficiency virus-positive women. *J Natl Cancer Inst* (2005) **97**:577–86. doi:10.1093/jnci/dji073
- Ho GY, Burk RD, Klein S, Kadish AS, Chang CJ, Palan P, et al. Persistent genital human papillomavirus infection as a risk factor for persistent cervical dysplasia. *J Natl Cancer Inst* (1995) **87**:1365–71. doi:10.1093/jnci/87.18.1365
- Lorincz AT, Castle PE, Sherman ME, Scott DR, Glass AG, Wacholder S, et al. Viral load of human papillomavirus and risk of CIN3 or cervical cancer. *Lancet* (2002) **360**:228–9. doi:10.1016/S0140-6736(02)09463-1
- Corden SA, Sant-Cassia LJ, Easton AJ, Morris AG. The integration of HPV-18 DNA in cervical carcinoma. *Mol Pathol* (1999) **52**:275–82. doi:10.1136/mp.52.5.275
- Arlt D, Huber W, Liebel U, Schmidt C, Majety M, Sauermann M, et al. Functional profiling: from microarrays via cell-based assays to novel tumor relevant modulators of the cell cycle. *Cancer Res* (2005) **65**:7733–42. doi:10.1158/0008-5472.CAN-05-0642
- April C, Klotzle B, Royce T, Wickham-Garcia E, Boyaniwsky T, Izzo J, et al. Whole-genome gene expression profiling of formalin-fixed, paraffin-embedded tissue samples. *PLoS One* (2009) **4**:e8162. doi:10.1371/journal.pone.0008162
- Waldron L, Ogino S, Hoshida Y, Shima K, McCart Reed AE, Simpson PT, et al. Expression profiling of archival tumors for long-term health studies. *Clin Cancer Res* (2012) **18**:6136–46. doi:10.1158/1078-0432.CCR-12-1915
- Delmas MC, Larsen C, Van Benthem B, Hamers FF, Bergeron C, Poveda JD, et al. Cervical squamous intraepithelial lesions in HIV-infected women: prevalence, incidence and regression. European study group on natural history of HIV infection in women. *AIDS* (2000) **14**:1775–84. doi:10.1097/00002030-200008180-00013
- Ellerbrock TV, Chiasson MA, Bush TJ, Sun XW, Sawo D, Brudney K, et al. Incidence of cervical squamous intraepithelial lesions in HIV-infected women. *JAMA* (2000) **283**:1031–7. doi:10.1001/jama.283.8.1031
- Xu C, Houck JR, Fan W, Wang P, Chen Y, Upton M, et al. Simultaneous isolation of DNA and RNA from the same cell population obtained by laser capture microdissection for genome and transcriptome profiling. *J Mol Diagn* (2008) **10**:129–34. doi:10.2353/jmoldx.2008.070131
- Chatterjee A, Stockwell PA, Rodger EJ, Morison IM. Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Res* (2012) **40**:e79. doi:10.1093/nar/gks150
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res* (2009) **19**:1639–45. doi:10.1101/gr.092759.109
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* (2012) **7**:562–78. doi:10.1038/nprot.2012.016
- Wang J, Duncan D, Shi Z, Zhang B. WEB-based gene set analysis toolkit (WebGestalt): update 2013. *Nucleic Acids Res* (2013) **41**:W77–83. doi:10.1093/nar/gkt439
- Batra J, O'Mara T, Patnala R, Lose F, Clements JA. Genetic polymorphisms in the human tissue kallikrein (KLK) locus and their implication in various malignant and non-malignant diseases. *Biol Chem* (2012) **393**:1365–90. doi:10.1515/hsz-2012-0211
- Sgroi DC, Teng S, Robinson G, Levangie R, Hudson JR Jr, Elkahoul AG. In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res* (1999) **59**:5656–61.
- Ma XJ, Salunga R, Tuggle JT, Gaudet J, Enright E, Mcquary P, et al. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci USA* (2003) **100**:5974–9. doi:10.1073/pnas.0931261100
- St Croix B, Rago C, Velculescu V, Traverso G, Romans KE, Montgomery E, et al. Genes expressed in human tumor endothelium. *Science* (2000) **289**:1197–202. doi:10.1126/science.289.5482.1197
- Porter PL. Molecular markers of tumor initiation and progression. *Curr Opin Genet Dev* (2001) **11**:60–3. doi:10.1016/S0959-437X(00)00157-X
- Polyak K, Riggins GJ. Gene discovery using the serial analysis of gene expression technique: implications for cancer research. *J Clin Oncol* (2001) **19**:2948–58.
- Luo JH, Yu YP, Cieply K, Lin F, DeFlavia P, Dhir R, et al. Gene expression analysis of prostate cancers. *Mol Carcinog* (2002) **33**:25–35. doi:10.1002/mc.10018
- Kitahara O, Furukawa Y, Tanaka T, Kihara C, Ono K, Yanagawa R, et al. Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia. *Cancer Res* (2001) **61**:3544–9.
- Birkenkamp-Demtroder K, Christensen LL, Olesen SH, Frederiksen CM, Laiho P, Aaltonen LA, et al. Gene expression in colorectal cancer. *Cancer Res* (2002) **62**:4352–63.
- Stremmel C, Wein A, Hohenberger W, Reingruber B. DNA microarrays: a new diagnostic tool and its implications in colorectal cancer. *Int J Colorectal Dis* (2002) **17**:131–6. doi:10.1007/s00384-001-0370-7
- Chang YE, Laimins LA. Microarray analysis identifies interferon-inducible genes and Stat-1 as major transcriptional targets of human papillomavirus type 31. *J Virol* (2000) **74**:4174–82. doi:10.1128/JVI.74.9.4174-4182.2000



32. Nees M, Geoghegan JM, Hyman T, Frank S, Miller L, Woodworth CD. Papillomavirus type 16 oncogenes downregulate expression of interferon-responsive genes and upregulate proliferation-associated and NF-kappaB-responsive genes in cervical keratinocytes. *J Virol* (2001) **75**:4283–96. doi:10.1128/JVI.75.9.4283-4296.2001
33. Ahn WS, Bae SM, Lee JM, Namkoong SE, Han SJ, Cho YL, et al. Searching for pathogenic gene functions to cervical cancer. *Gynecol Oncol* (2004) **93**:41–8. doi:10.1016/j.ygyno.2003.11.031
34. Paliouras M, Borgono C, Diamandis EP. Human tissue kallikreins: the cancer biomarker family. *Cancer Lett* (2007) **249**:61–79. doi:10.1016/j.canlet.2006.12.018
35. Anisowicz A, Sotiropoulou G, Stenman G, Mok SC, Sager R. A novel protease homolog differentially expressed in breast and ovarian cancer. *Mol Med* (1996) **2**:624–36.
36. Ogawa K, Utsunomiya T, Mimori K, Tanaka F, Inoue H, Nagahara H, et al. Clinical significance of human kallikrein gene 6 messenger RNA expression in colorectal cancer. *Clin Cancer Res* (2005) **11**:2889–93. doi:10.1158/1078-0432.CCR-04-2281
37. Santin AD, Diamandis EP, Bellone S, Soosaipillai A, Cane S, Palmieri M, et al. Human kallikrein 6: a new potential serum biomarker for uterine serous papillary cancer. *Clin Cancer Res* (2005) **11**:3320–5. doi:10.1158/1078-0432.CCR-04-2528
38. Heizmann CW, Fritz G, Schafer BW. S100 proteins: structure, functions and pathology. *Front Biosci* (2002) **7**:d1356–68. doi:10.2741/heizmann
39. Glaser R, Koten B, Wittersheim M, Harder J. Psoriasin: key molecule of the cutaneous barrier? *J Dtsch Dermatol Ges* (2011) **9**:897–902. doi:10.1111/j.1610-0387.2011.07683.x
40. Kaur J, Matta A, Kak I, Srivastava G, Assi J, Leong I, et al. S100A7 overexpression is a predictive marker for high risk of malignant transformation in oral dysplasia. *Int J Cancer* (2014) **134**:1379–88. doi:10.1002/ijc.28473
41. Hattinger E, Zwicker S, Ruzicka T, Yuspa SH, Wolf R. Opposing functions of psoriasin (S100A7) and koebnerisin (S100A15) in epithelial carcinogenesis. *Curr Opin Pharmacol* (2013) **13**:588–94. doi:10.1016/j.coph.2013.04.007
42. Rajkumar T, Sabitha K, Vijayalakshmi N, Shirley S, Bose MV, Gopal G, et al. Identification and validation of genes involved in cervical tumorigenesis. *BMC Cancer* (2011) **11**:80. doi:10.1186/1471-2407-11-80
43. Peng X, Wu Z, Yu L, Li J, Xu W, Chan HC, et al. Overexpression of cystic fibrosis transmembrane conductance regulator (CFTR) is associated with human cervical cancer malignancy, progression and prognosis. *Gynecol Oncol* (2012) **125**:470–6. doi:10.1016/j.ygyno.2012.02.015
44. Gupta N, Prasad PD, Ghamande S, Moore-Martin P, Herdman AV, Martindale RG, et al. Up-regulation of the amino acid transporter ATB(0,+)(SLC6A14) in carcinoma of the cervix. *Gynecol Oncol* (2006) **100**:8–13. doi:10.1016/j.ygyno.2005.08.016
45. Chen J, Yao Y, Gong C, Yu F, Su S, Chen J, et al. CCL18 from tumor-associated macrophages promotes breast cancer metastasis via PTPN23. *Cancer Cell* (2011) **19**:541–55. doi:10.1016/j.ccr.2011.02.006
46. Gnanaprasadam VJ. Unlocking the molecular archive: the emerging use of formalin-fixed paraffin-embedded tissue for biomarker research in urological cancer. *BJU Int* (2010) **105**:274–8. doi:10.1111/j.1464-410X.2009.08665.x
47. Scicchitano MS, Dalmas DA, Bertiaux MA, Anderson SM, Turner LR, Thomas RA, et al. Preliminary comparison of quantity, quality, and microarray performance of RNA extracted from formalin-fixed, paraffin-embedded, and unfixed frozen tissue samples. *J Histochem Cytochem* (2006) **54**:1229–37. doi:10.1369/jhc.6A6999.2006
48. Penland SK, Keku TO, Torrice C, He X, Krishnamurthy J, Hoadley KA, et al. RNA expression analysis of formalin-fixed paraffin-embedded tumors. *Lab Invest* (2007) **87**:383–91.
49. Frank M, Doring C, Metzler D, Eckerle S, Hansmann ML. Global gene expression profiling of formalin-fixed paraffin-embedded tumor samples: a comparison to snap-frozen material using oligonucleotide microarrays. *Virchows Arch* (2007) **450**:699–711. doi:10.1007/s00428-007-0412-9
50. Coudry RA, Meireles SI, Stoyanova R, Cooper HS, Carpino A, Wang X, et al. Successful application of microarray technology to microdissected formalin-fixed, paraffin-embedded tissue. *J Mol Diagn* (2007) **9**:70–9. doi:10.2353/jmoldx.2007.060004

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 08 September 2014; accepted: 11 November 2014; published online: 26 November 2014.

Citation: Royse KE, Zhi D, Conner MG, Clodfelder-Miller B, Srinivasainagendra V, Vaughan LK, Skibola CF, Crossman DK, Levy S and Shrestha S (2014) Differential gene expression landscape of co-existing cervical pre-cancer lesions using RNA-seq. *Front. Oncol.* **4**:339. doi: 10.3389/fonc.2014.00339

This article was submitted to *Cancer Genetics*, a section of the journal *Frontiers in Oncology*.

Copyright © 2014 Royse, Zhi, Conner, Clodfelder-Miller, Srinivasainagendra, Vaughan, Skibola, Crossman, Levy and Shrestha. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.